AWARD NUMBER: W81XWH-14-1-0234

TITLE: Single-Cell RNA Sequencing of the Bronchial Epithelium in Smokers With Lung Cancer

PRINCIPAL INVESTIGATOR: Jennifer Beane-Ebel

CONTRACTING ORGANIZATION: Boston University School of Medicine
Boston, MA 02118

REPORT DATE: July 2015

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED**
July 2015	Annual	1 Jul 2014 - 30 Jun 2015

4. TITLE AND SUBTITLE
Single-Cell RNA Sequencing of the Bronchial Epithelium in Smokers With Lung Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-14-1-0234

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Jennifer Beane-Ebel, Joshua Campbell, Grant Duclos

E-Mail: jbeane@bu.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Boston University School of Medicine
Department of Medicine
Division of Computational Biomedicine
Medical Campus
72 East Concord Street, E-631
Boston, MA 02118-2308

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Cigarette smoking, the major cause of lung cancer, creates a "field of injury" throughout the respiratory tract. We have previously shown that gene expression from bronchial epithelial cells reflects the physiologic response to cigarette smoke exposure and can serve as a diagnostic biomarker for lung cancer. The purpose of this Idea Development Award is to conduct single cell RNA sequencing on airway epithelial cells obtained from smokers with and without lung cancer to identify cell-type dependent gene expression alterations in the lung cancer field of injury.
Cells are being collected by brushing the right mainstem bronchus of smokers undergoing bronchoscopy for the suspicion of lung cancer. A protocol has been developed to isolate single cells from these bronchial brushings using fluorescence-activated cell sorting (FACS). Next, we have implemented an adapted version of the CEL-Seq RNA library preparation protocol that includes plate-, well-, and transcript-specific barcodes allowing hundreds of cells to be pooled together and sequenced. We have also developed a computational pipeline to process the sequencing data into gene level counts for each cell. In order to demonstrate that the methodologies described are working, a pilot experiment was conducted in which bronchial epithelial cells from 1 former smoker and 1 current smoker were profiled (n=24 cells per donor). Data generated from this experiment illustrated that there is low cell-to-cell technical variation and known smoking-associated gene expression alterations can be detected. The success of this initial experiment is significant because it demonstrates that all the protocols are currently in place to begin processing samples collected from smokers with and without lung cancer.
Over the next year we plan to sequence hundreds of cells per donor from about 30 smokers with and without lung cancer to discover known and novel cell types with gene expression changes associated with the presence of lung cancer. These discoveries may enhance current lung cancer diagnostics as well as suggest potential new therapeutics for lung cancer.

15. SUBJECT TERMS
single cell, bronchial epithelium, mRNA sequencing, and lung cancer

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			USAMRMC
Unclassified	Unclassified	Unclassified	Unclassified	12	**19b. TELEPHONE NUMBER** *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

ABSTRACT:
Cigarette smoking, the major cause of lung cancer, creates a "field of injury" throughout the respiratory tract. We have previously shown that gene expression from bronchial epithelial cells reflects the physiologic response to cigarette smoke exposure and can serve as a diagnostic biomarker for lung cancer. The purpose of this Idea Development Award is to conduct single cell RNA sequencing on airway epithelial cells obtained from smokers with and without lung cancer to identify cell-type dependent gene expression alterations in the lung cancer field of injury.

Cells are being collected by brushing the right mainstem bronchus of smokers undergoing bronchoscopy for the suspicion of lung cancer. A protocol has been developed to isolate single cells from these bronchial brushings using fluorescence-activated cell sorting (FACS). Next, we have implemented an adapted version of the CEL-Seq RNA library preparation protocol that includes plate-, well-, and transcript-specific barcodes allowing hundreds of cells to be pooled together and sequenced. We have also developed a computational pipeline to process the sequencing data into gene level counts for each cell. In order to demonstrate that the methodologies described are working, a pilot experiment was conducted in which bronchial epithelial cells from 1 former smoker and 1 current smoker were profiled (n=24 cells per donor). Data generated from this experiment illustrated that there is low cell-to-cell technical variation and known smoking-associated gene expression alterations can be detected. The success of this initial experiment is significant because it demonstrates that all the protocols are currently in place to begin processing samples collected from smokers with and without lung cancer.

Over the next year we plan to sequence hundreds of cells per donor from about 30 smokers with and without lung cancer to discover known and novel cell types with gene expression changes associated with the presence of lung cancer. These discoveries may enhance current lung cancer diagnostics as well as suggest potential new therapeutics for lung cancer.

INTRODUCTION:
Cigarette smoking, the major cause of lung cancer, creates a "field of injury" throughout the respiratory tract by inducing molecular alterations such as allelic loss, p53 mutations, changes in promoter methylation and telomerase activity[1-5]. We have previously shown that gene expression from bronchial epithelial cells reflects the physiologic response to cigarette smoke exposure[6,7]. Importantly, we have extended this airway field of injury to the study of lung cancer, and have identified a bronchial airway gene expression signature that can serve as a diagnostic biomarker for lung cancer[8] that performs independently of clinical risk factors for disease[9]. The lung cancer diagnostic biomarker has been subsequently validated in in a large clinical trial[10] and has been commercialized by Veracyte, Inc. and is known as PERCEPTA™. Advances in technology for amplification of low amounts of RNA combined with next-generation sequencing have produced the ability to characterize the transcriptome of individual cells. While the bronchial brushings examined in our previous studies have captured a relatively pure population of bronchial epithelial cells, we are unable to discern which airway cell type or types are responsible for the gene expression changes observed nor characterize gene expression variation between cells. Variation in gene expression across single cells can be used to define unique subpopulations of cells that may be independent of known markers or cell morphology that may associate with lung cancer. We hypothesize that the lung cancer-specific gene expression in the bronchial epithelium might be restricted to specific known cell types (e.g. basal cells) or molecularly defined subpopulations of cells. The goal of this study will be to use single-cell RNA sequencing to identify cell-type dependent gene expression alterations in the lung cancer field of injury and to molecularly identify novel subpopulations of cells that are associated with lung cancer. These novel molecular insights hold the potential to improve the diagnostic utility of the airway epithelium for lung cancer and to guide new therapeutic strategies for lung cancer prevention.

KEYWORDS: single cell, bronchial epithelium, mRNA sequencing, and lung cancer

ACCOMPLISHMENTS:
What were the major goals of the project?

- Specific Aim 1: Identify which cell types in the airway epithelium harbor the lung cancer-specific alterations in biomarker genes by single cell RNA sequencing
    - Major Task 1: Isolate and sequence the RNA of single epithelial cells from the bronchus of smokers with and without lung cancer (n=15 subjects/group, n=960 single cells/subject).
        - Subtask 1: Approval of IRB and HRPO (1-2)

          The IRB at Boston University School of Medicine approved the study protocol on September 10, 2014 but notification of the outcome was received on December 4, 2014. The HRPO approval was obtained on December 19, 2014. This process took approximately 6 months to complete and delayed sample collection by about 4 months. Completion percent: 100%

- Subtask 2: Collection of airway brushings from 30 subjects at BUMC (2-12)

  We have collected airway brushings from 9 subjects at BUMC in the past 6 months after full study approval. We have collected 9 brushings from current and former smokers under clinical suspicion of lung cancer. Of those subjects, 5 have been confirmed to have lung cancer, 3 do not have lung cancer, and 1 has yet to be determined (Table 1). Additionally, from an external cohort (see "Sample Collection" section under "What was accomplished under these goals?" we have collected 11 brushings from healthy current smokers and 10 brushings from healthy former smokers.

  As collection has been slower than expected we are in the process of collaborating with Dr. Robert Browning, an interventional pulmonologist at Walter Reed Medical Center, to collect additional brushings. We are currently adding his center to our IRB protocol. Completion Percent: 33%

- Subtask 3: Sorting of cells from brushings using FACS (Sorting of cells will take place within hours after collection) (2-12)

  Tissue acquired via bronchial brushing is dissociated, dead cells and red blood cells are excluded via FACS, and live cells are sorted into 96-well plates and frozen. This process needs to be completed immediately after sample collection. Therefore, we have sorted all samples collected above and completion is dependent on sample collection.
  Completion Percent: 33%

- Subtask 4: mRNA isolation and library preparation (2-13)

  An established technique (CEL-Seq) for preparing single cell RNA-Seq libraries was adapted for this project and modified to increase sample multiplexing capacities and correct for experimental amplification biases. To date, we have produced high quality single cell data using this protocol and in the coming year we will process the samples collected. Completion Percent: 33%

- Subtask 5: Sequencing of samples on Ilumina HiSeq 2500 (12-14)
  Single cell RNA libraries are massively multiplexed and paired-end sequencing is performed using the Illumina HiSeq 2500. Currently we have only sequenced cells from our pilot experiments. Completion Percent: 15%

- Subtask 6: De-multiplex samples, preprocess, align, and analyze data quality (12-18)

  A computational pipeline developed in collaboration with the Yanai Lab (http://yanailab.technion.ac.il/) was used to preprocess and align reads generated via the CEL-Seq methodology. Additional metrics have been incorporated for the purposes of quality control and determination of sample cell type. The pipeline developed will be used to process the data as it is generated. Percent complete: 25%

- Milestone(s) Achieved: Generation of high quality single cell RNA sequencing data from 30 subjects

  As stated above, due to delayed IRB approval we didn't start collect samples when we expected. In order to quickly collect the numbers of samples proposed we are adding another collection site. Once the samples are collected, all the protocols are now in place to process them. We expect completion of this milestone within the next year.

o Major Task 2: Determine the cell type(s) responsible for the aberrant gene expression in the airway lung cancer diagnostic biomarker. Completion Percent: 0%
  - Subtask 1: Summarize sequencing data into counts per gene (12-18)
  - Subtask 2: Classify gene expression as signal or noise based on a mixture model (12-18)
  - Subtask 3: Determine cell types of origin for lung cancer biomarker genes (12-20)
  - Subtask 4: Identify cell type dependent lung cancer associated differential expression using a linear modeling and ANOVA strategy (18-22)

        - Subtask 5: Choose candidates for validation (22)
        - Milestone(s) Achieved: Identification of which cell types in the airway epithelium express lung-cancer specific gene expression alterations

- Specific Aim 2: Identify unique cell populations in the airway of smokers that are associated with lung cancer. Completion Percent: 0%
    - Major Task 1: Molecularly identify subpopulations of cells irrespective of cell type and determine if these cells are more or less abundant in the airways of patients with lung cancer
        - Subtask 1: Identification of novel subpopulations of cells using both class discovery and pathway prediction approaches (18-22)
        - Subtask 2: Identify subpopulations associated with lung cancer
        - Subtask 3: Choose candidates for validation (18-22)
        - Milestone(s) Achieved: Identification of novel subpopulations of airway epithelial cells that are associated with lung cancer (22)
    - Major Task 2: Validate lung cancer associated genes from specific cell types or from novel subpopulations in bronchial epithelial cells from independent subjects (n=10) using FISH
        - Subtask 1: Collect airway brushes from 10 subject for validation (13-22)
        - Subtask 2: RNA-FISH will be used to validate 5 candidate genes in conjunction with 4 known epithelial marker genes (22-24)
        - Milestone(s) Achieved: Validation of novel lung cancer-associated gene candidates in specific populations of epithelial cells

What was accomplished under these goals?
The major objective is to conduct single cell RNA sequencing on cells collected from bronchial brushings from current and former smokers with and without lung cancer. In order to attain this goal several important experimental protocols needed to be developed and validated. We have developed an unbiased methodology for sorting the cells from bronchial brushings into 96-well plates, a library preparation protocol that will produce high quality data, and an analysis pipeline for processing the data (Figure 1). In order to accomplish these tasks we have and are continuing to conduct sequencing experiments to validate that our methods are working prior to processing brushings from smokers with and without lung cancer.

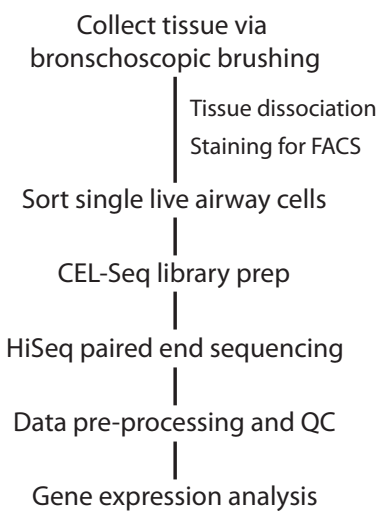

Figure 1. Flow diagram depicting steps involved in generation and analysis of single cell RNA sequencing data from bronchial airway cells.

Sample Collection
Subjects recruited are current and former smokers undergoing flexible bronchoscopy for clinical suspicion of lung cancer at BUMC. For each consented subject, we collect data regarding their age, gender, race, and a detailed smoking history (Table 1).

Additional samples were collected from healthy current and former smoker volunteers recruited for a project entitled "Airway Epithelium Profiling for Evaluation of E-cigarettes & Tobacco Products" (Table 1). The subjects in this study were recruited via advertisements, a two-part screening questionnaire was administered via telephone to determine eligibility, and eligible subjects completed two study visits, the second of which included a bronchoscopy. These healthy donors will be treated as controls and data generated for the "Airway Epithelium Profiling for Evaluation of E-cigarettes & Tobacco Products" project will be used in conjunction with data generated from samples collected for this project.

	Clinical Suspicion of Lung Cancer	Healthy Volunteer
Cancer Diagnosis	5 Cancer, 3 No Cancer, 1 TBD	21 No Cancer
Smoking Status	5 Current, 4 Former	11 Current, 10 Former
Age	65 (5)	44 (11)
Pack years (SD)	38 (34)	11 (16)
Sex	3 Male, 6 Female	10 Male, 11 Female

Table 1. Demographic and clinical information from bronchial brushing donors.

Collection of bronchial brushings is done using the same technique as in our prior studies[6,8]. Following topical anesthesia of the upper airway using 2% lidocaine, a bronchoscope is introduced to the right mainstem bronchus and epithelial cells are obtained using an endoscopic cytobrush.

Cell sorting by FACS

Tissue obtained from bronchial brushings is treated with 0.25% Trypsin/EDTA for epithelial sheet dissociation and cells are sorted using a BD FACSAria II. Gating based on forward scatter height vs. forward scatter area (FSC-H vs. FSC-A) is applied to sort only singlet events (single cells). Staining for GYPA (CD235a) is used to exclude all red bloods cells. Hoechst 33342 is used to stain the DNA of all cells, whereas PI is used to specifically stain the DNA of dead cells with compromised membranes. For each donor, single Hoechst 33342^{+} PI^{-} CD235a^{-} cells are sorted into five 96-well PCR plates (480 cells), frozen on dry ice and stored at -80 °C until preparation for sequencing (Figure 2).

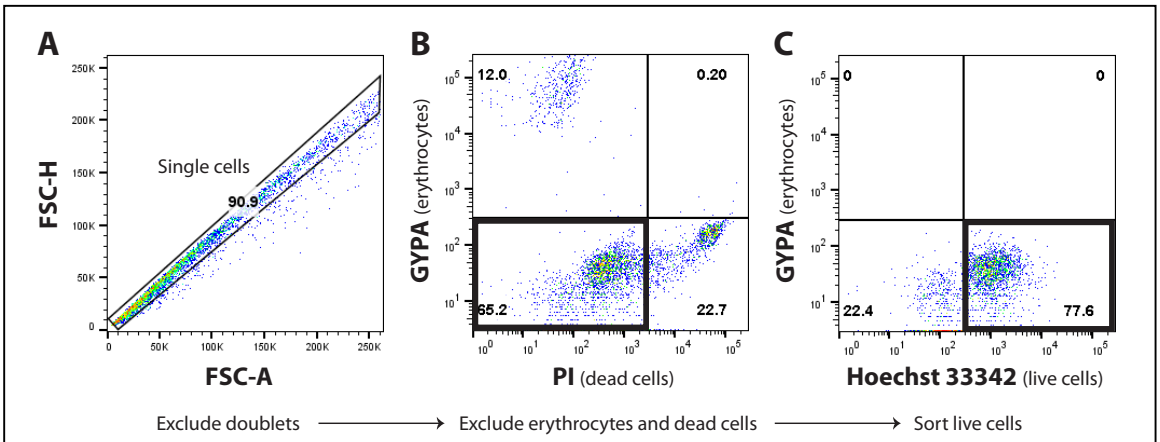

Figure 2. Sorting of bronchial airway cells. FACS is used to **(A)** isolate singlet events based on forward scatter height vs. forward scatter area (FSH-H vs. FSH-A). **(B)** Dead cells (PI^{+}) and red blood cells (GYPA^{+}) are stained and excluded, and **(C)** live cells (Hoechst 33342^{+}) are sorted.

Development of a single cell mRNA sequencing protocol

Massively parallel single cell RNA-sequencing of human bronchial airway cells is being performed using the CEL-Seq RNA library preparation protocol that has been modified for increased high throughput capacities[11]. Frozen 96-well PCR plates containing sorted cells are thawed on ice and nucleic acids are purified from cell lysates using RNAClean XP beads. RNA is reverse transcribed (Ambion AM1751) using primers composed of an anchored poly(dT), the 5' Illumina adaptor sequence, a well-specific barcode, a random sequence, which serves as a transcript-specific unique molecular identifier (UMI)[12], and a T7 RNA polymerase promoter. Samples are additionally supplemented with ERCC RNA Spike-In mix (Ambion) for quality control[13]. cDNA generated from each of the 96 cells per plate is uniquely barcoded and therefore can be pooled for second strand synthesis (Ambion AM1751) and amplification by in vitro transcription (Ambion AM1751). Amplified RNA is then chemically fragmented (NEB E6150) and ligated to the Illumina RNA 3' adapter (Illumina RS-200-0012). Samples are again reverse transcribed and amplified using indexed Illumina RNA PCR primers (Illumina RS-200-0012). This barcoding strategy allows for the loading of libraries derived up to 4608 cells (96 well-specific barcodes x 48 plate-specific indices) onto a single flow cell lane for paired-end sequencing using the Illumina HiSeq 2500.

Development of a computational pipeline to process sequencing data

In collaboration with the Yanai lab (http://yanailab.technion.ac.il/), we utilized a computational pipeline to preprocess and align reads generated from the CEL-Seq protocol. Briefly, reads were demultiplexed into "plate" level FASTQ files using Illumina software. Each plate FASTQ file was further demultiplexed into cell specific FASTQ files using custom scripts. The UMI barcode was also removed from each read and stored in the read header. Reads were aligned to the human genome using Bowtie2. Gene level counts were derived using a modified version of the HTSeq python library, which only counts multiple reads that are aligned to the same position and have the same UMI once, thus eliminating PCR amplification bias. Quality of each cell was assessed by examining the total number of reads aligned, the total number of genes detected, the number of ERCC spike in transcripts detected, and known airway cell type markers.

Pilot Sequencing Experiment

A pilot experiment was designed to determine if the described single cell RNA sequencing approach can be used to effectively detect a response to cigarette smoke exposure at single cell resolution and to assess whether or not this method generates significant cell-to-cell technical variation. Bronchial epithelial cells from 1 healthy former smoker and 1 healthy current smoker (n=24 cells per donor) were profiled (48 total cells) via paired end sequencing using the Illumina MiSeq. Epithelial cells were sorted using the FACSAria II, according to expression of the established airway epithelial surface marker CD166[14]. Expression of major epithelial subpopulation markers was detected in subsets of cells profiled (KRT5, MUC5AC, SCGB1A1, FOXJ1). A surface marker was used for cell sorting in this pilot experiment to specifically generate RNA sequencing data from airway epithelial cells, but all future experiments involve unbiased sorting of live cells to profile all cellular subpopulations present within bronchial brushings.

To assess technical variation of the sequencing protocol, the ERCC RNA Spike-In Mix (Ambion 4456740) is added to RNA from each individual cell prior to beginning CEL-Seq library preparation. The ERCC RNA Spike-In Mix is a set of RNA controls added at a range of known concentrations. The mean detection of ERCC transcripts across all 48 cells strongly correlates (Pearson's r = 0.92) with known concentrations (Fig. 3A). Results indicate that this sequencing approach can be used to generate data with negligible cell-to-cell technical variation.

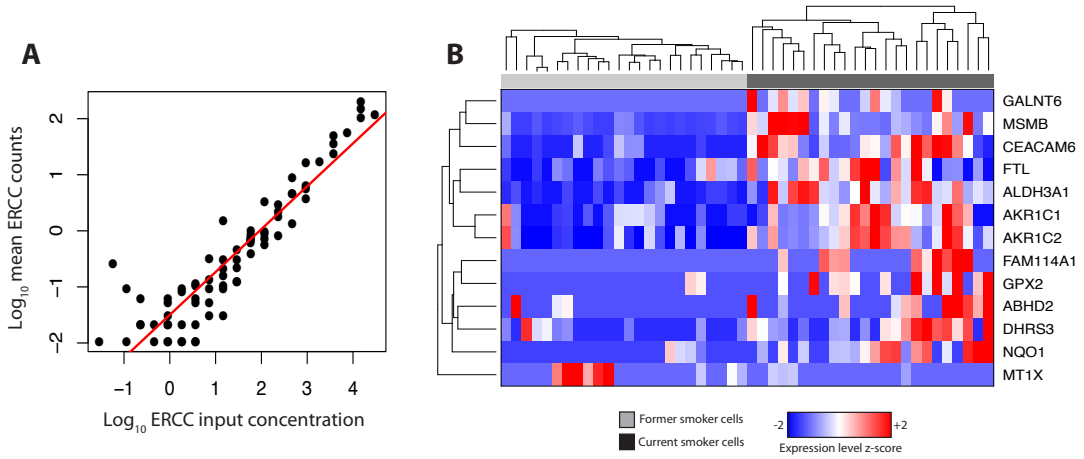

Figure 3. RNA-Seq profiling of single bronchial epithelial cells from healthy former and current smokers. Single bronchial epithelial cells from 1 former smoker and 1 current smoker (n=24 cells each) were profiled using the CEL-Seq protocol. **(A)** Mean detection levels of ERCC RNA Spike-In transcripts across all 48 cells strongly correlated (Pearson's r = 0.92) with known concentrations of each transcript added to each sample. **(B)** Differentially expressed genes (FDR $q < 0.05$) associated with smoking status overlapped with and are changing concordantly with 13 genes previously reported to change in the airways of current smokers [ref = Beane et al. 2007].

Negative binomial generalized linear modeling was used to identify genes whose expression changed with respect to smoking status (FDR q < 0.05). Differentially expressed genes were intersected with a set of genes previously reported to change in the airways of current and former smokers (Fig. 3B)[7]. All overlapping genes (n=13) directionally changed in concordance with previously reported findings. Additional donors and increased numbers of cells per donor are necessary to properly assess the single cell resolution response to smoking, but these preliminary findings illustrate that it is likely that single cell RNA sequencing will allow us to understand the cell type-specific contributions to the lung cancer field of injury.

What opportunities for training and professional development has the project provided?
Grant Duclos, a graduate student, responsible for sorting the cells and preparing the libraries as part of the project has had the opportunity to study under Dr. Itai Yanai, Associate Professor at Technion – Israel Institute of Technology. Dr. Yanai is on sabbatical at the Broad Institute and has extensive experience in single cell sequencing and under his mentorship Grant has been able to effectively troubleshoot our protocols.

How were the results disseminated to communities of interest?
Nothing to Report.

What do you plan to do during the next reporting period to accomplish the goals?
In order to accomplish the goals in the Statement of Work, over the next year we plan to complete our sample collection. We will accomplish this by continuing to collect samples from subjects undergoing bronchoscopy for suspicion of lung cancer at Boston University Medical Center. We also plan to collect additional samples through collaboration with Dr. Robert Browning, an interventional pulmonologist at Walter Reed National Military Medical Center. Grant Duclos from our group will travel to Maryland to train Dr. Browning's group to collect, sort, and freeze cells obtained at bronchoscopies that will be shipped to BU for processing. Currently, we are in the process of amending our IRB to include this additional site. Additionally, we have shown above that we have successfully created protocols/methods for collecting, sorting, processing, and analyzing the data; therefore, within the next year we plan to generate data on approximately 200 cells from about 30 smokers with and without lung cancer. We will analyze the data to discover cell type dependent lung cancer-associated gene expression alterations in known and unique cell populations. In addition, we will continue to collect samples for RNA-FISH validation.

IMPACT:
What was the impact on the development of the principal discipline(s) of the project?
To date, human bronchial epithelial cells obtained via bronchoscopy have not been sorted and processed for single cell RNA sequencing. The cells need to be immediately taken from the bronchoscopy suite and processed for FACS sorting.

FACS sorting of the cells is accomplished as stated above and the sorted cells are frozen in 96-well plates. The CEL-Seq RNA preparation protocol has been modified to provide the ability to process hundreds if cells from several subjects. All of the methodology and protocols developed as part of this project will directly benefit other groups that are attempting to examine single cell transcriptomics in human clinical samples. In addition, all microarray or RNA sequencing studies to date on human airway epithelial cells have been conducted using a bulk population of cells. Expression values represent the mean behavior of all the cells profiled in a given sample and do not capture cell-to-cell gene expression variation. Using a single cell approach, we will be able to characterize the cell types (both known and novel) that are present within the subjects and identify the cell types responsible for the lung cancer-specific airway field of injury. Using our sorting technique, we will also be profiling immune cells present when the brushing was obtained. Profiling of both epithelial and immune cells will allow us to look at the interaction between the cell types and begin to characterize the impact of the immune system in maintaining a healthy epithelium. The study will contribute to our understanding of the human airway epithelium and the changes that occur in response to smoke exposure and the acquisition of lung cancer.

What was the impact on other disciplines?
The project is multidisciplinary and is likely to impact the study of lung cancer and epithelial cell biology as well as contribute to molecular biology and bioinformatics methods. An example of a broad impact would be if the study suggests new markers to study rare epithelial cell types that appear to be important in the process of lung carcinogenesis. Another example would be if bioinformatics techniques developed during the analysis of the data for the project could be applied to similar datasets.

.

What was the impact on technology transfer?
Nothing to report.

What was the impact on society beyond science and technology?
We have developed a gene expression based biomarker for improving lung cancer diagnosis known as PERCEPTA™ that has been commercialized by Vercyte (http://www.veracyte.com). The test helps identify patients at low risk for having lung cancer after a non-diagnostic bronchoscopy ordered as a result of CT scan abnormalities. The findings in this project may help establish a more sensitive diagnostic test that will impact the clinical management of high lung risk patients.

CHANGES/PROBLEMS:
Changes in approach and reasons for change
We have previously gained approval to switch single cell sequencing protocols from SMART-Seq to CEL-Seq in order to increase the total number of cells that will be profiled per patient. Using this protocol allowed us to switch to a new sorting procedure that captures all epithelial, immune, and potentially unknown cell types in an unbiased fashion. We have also expanded sample collection to include bronchial brushings collected from healthy current and former smoker volunteers recruited for a project entitled “Airway Epithelium Profiling for Evaluation of E-cigarettes & Tobacco Products”. These healthy donors will be treated as controls and data generated for the “Airway Epithelium Profiling for Evaluation of E-cigarettes & Tobacco Products” project will be used in conjunction with data generated from samples collected for this project. Determining which cell types the lung cancer biomarker genes are expressed in these additional control groups will increase our understanding of the nature of the “field of injury” and why these genes are differentially expressed in lung cancer patients only.

Actual or anticipated problems or delays and actions or plans to resolve them
With CEL-Seq, we are detecting ~2000 genes per cell on average. It is possible that this coverage will not be enough to distinguish all subpopulations of cells within the airway brushings. If this occurs, we will either perform deeper sequencing of the CEL-Seq libraries to detect more rare transcripts or perform SMART-Seq on an additional set of cells which is reported to detect ~5000 genes on average[15].

In addition, we have had delays in sample collection because of the IRB approval process. We will continue collecting samples from Boston University Medical Center and will also collect brushes in collaboration with Dr. Robert Browning’s group at Walter Reed Medical Center as explained above.

Changes that had a significant impact on expenditures
Nothing to Report.

Significant changes in use or care of human subjects, vertebrate animals, biohazards, and/or select agents
Nothing to report.

Significant changes in use or care of human subjects
Nothing to report.

Significant changes in use or care of vertebrate animals.
Nothing to report.

Significant changes in use of biohazards and/or select agents
Nothing to report.

PRODUCTS:
Publications, conference papers, and presentations
Nothing to report.
Journal publications.
Nothing to report.

Books or other non-periodical, one-time publications.
Nothing to report.

Other publications, conference papers, and presentations.
Nothing to report.

Website(s) or other Internet site(s)
Nothing to report.

Technologies or techniques
The techniques, including cell sorting, library preparation, and analysis methods developed in this project will be shared through publication of a manuscript reporting the findings of single cell sequencing experiments profiling smokers with the without lung cancer.

Inventions, patent applications, and/or licenses
Nothing to report.

Other Products
Nothing to report.

PARTICIPANTS & OTHER COLLABORATING ORGANIZATIONS
What individuals have worked on the project?

Name:	*Jennifer Beane-Ebel*
Project Role:	*Principal Investigator*
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	*3*
Contribution to Project:	Overseeing all aspects of the project including sample collection, experimental design, and data analysis
Funding Support:	DOD IDA, LUNGevity Foundation, NIH/NIAID, Industry award, Internal funds

Name:	*Joshua Campbell*
Project Role:	Postdoctoral Fellow
Researcher Identifier (e.g.	

ORCID ID):	
Nearest person month worked:	5
Contribution to Project:	Overseeing all aspects of the project with the PI and leading the computational analysis of the data
Funding Support:	DOD IDA, NIH/NHLBI, Industry funds

Name:	*Martine Dumas*
Project Role:	Study Coordinator
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	1
Contribution to Project:	Consenting patients and sample collection
Funding Support:	DOD IDA, NIH/NCI, Industry funds

Name:	*Grant Duclos*
Project Role:	Graduate Student
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	6
Contribution to Project:	Collecting samples, sorting cells, and preparing RNA sequencing libraries from the cells, and data analysis
Funding Support:	DOD IDA, NIH/NHLBI

Name:	*Yaron Gesthalter*
Project Role:	
Researcher Identifier (e.g. ORCID ID):	
Nearest person month worked:	NIH/NI
Contribution to Project:	Consenting patients and collection of samples
Funding Support:	NIH/NHLBI, Internal funds

Has there been a change in the active other support of the PD/PI(s) or senior/key personnel since the last reporting period?
Nothing to Report

What other organizations were involved as partners?
Organization Name: Broad Institute
Location of Organization: Boston, MA
Partner's contribution to the project: Collaboration. We have been collaborating with Itai Yanai, Associate Professor at Technion – Israel Institute of Technology during his sabbatical at the Broad Institute to help develop our single cell sequencing library preparation protocol.

SPECIAL REPORTING REQUIREMENTS
None.

APPENDICES:

References

1. Franklin, W. A. *et al.* Widely dispersed p53 mutation in respiratory epithelium. A novel mechanism for field carcinogenesis. *J. Clin. Invest.* **100,** 2133–2137 (1997).
2. Wistuba, I. I. *et al.* Molecular damage in the bronchial epithelium of current and former smokers. *J. Natl. Cancer Inst.* **89,** 1366–1373 (1997).
3. Tang, X. *et al.* EGFR tyrosine kinase domain mutations are detected in histologically normal respiratory epithelium in lung cancer patients. *Cancer Res.* **65,** 7568–7572 (2005).
4. Mao, L. *et al.* Clonal genetic alterations in the lungs of current and former smokers. *J. Natl. Cancer Inst.* **89,** 857–862 (1997).
5. Powell, C. A., Klares, S., O'Connor, G. & Brody, J. S. Loss of heterozygosity in epithelial cells obtained by bronchial brushing: clinical utility in lung cancer. *Clin. Cancer Res.* **5,** 2025–2034 (1999).
6. Spira, A. *et al.* Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 10143–10148 (2004).
7. Beane, J. *et al.* Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. *Genome Biol.* **8,** R201 (2007).
8. Spira, A. *et al.* Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat. Med.* **13,** 361–366 (2007).
9. Beane, J. *et al.* A prediction model for lung cancer diagnosis that integrates genomic and clinical features. *Cancer Prev Res (Phila)* **1,** 56–64 (2008).
10. Silvestri, G. A. *et al.* A Bronchial Genomic Classifier for the Diagnostic Evaluation of Lung Cancer. *N. Engl. J. Med.* (2015). doi:10.1056/NEJMoa1504601
11. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep* **2,** 666–673 (2012).
12. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat Meth* **9,** 72–74 (2012).
13. Baker, S. C. *et al.* The External RNA Controls Consortium: a progress report. *Nat Meth* **2,** 731–734 (2005).
14. Hegab, A. E. *et al.* Isolation and in vitro characterization of Basal and submucosal gland duct stem/progenitor cells from human proximal airways. *Stem Cells Transl Med* **1,** 719–724 (2012).
15. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344,** 1396–1401 (2014).